A Formal Study on Summary Table Handling with Application to Health Care Statistical Databases

Mihoko Okada, Ph.D., Minoru Takaba
Department of Medical Information Science
Suzuka University of Medical Science and Technology
Suzuka City, Mie, Japan

There is an ever-increasing demand for statistics in the health care field. Various surveys are conducted every year, and a huge number of summary tables are created. To share information and to make the most of the summary tables, we present a method for storing and transforming summary tables with special application to the health care field. A database management system being developed is also presented.

INTRODUCTION

In the field of health care, there is an ever-increasing demand for statistical information including statistics on diseases, mortality rates, patient billing, etc., and various surveys are conducted every year. As a result of a survey, a number of summary tables are created. In a national survey on clinical institutions, for example, typical examples include a table showing the number of hospitals classified by locations and by bed-size, a table showing the number of inpatients and length of stay classified by the types of hospitals, and so on. Such tables are distributed as publications (sometimes magnetic tape files are also available) conventionally. When one is interested in the survey results, it is not allowed to access the original data, and the summary tables are considered to be the most important source of information which can be shared. To make the most of the summary tables created year by year, a well-organized database should be provided.

Databases designed especially for providing statistics on groups of individuals are called statistical databases (SDBs). Many research works have focused on this special class of databases in the past. As to data model, Chan, et al proposed SUBJECT[1] for organizing and accessing large SDBs. Sato, et al proposed a model called SDM4S which is an extension of the relational model[2]. As to application systems, recent topics include the problem of user interface[3]. In the field of health care, there are a number of papers related to the statistical analyses on databases, but only few discuss the aspect of statistical data management[4]. We have been developing an SDB management system with special application to the health care field. In our study, an SDB is defined as a collection of summary tables. In a given survey, any number of tables can be produced conceptually, but in reality, only a limited number of tables which are considered important by the people concerned are created. From existent tables, however, more tables can be produced. Our study objective is to establish a method for storing summary tables systematically and provide an easy means to retrieve more tables than are physically stored. In this paper, we define the operations of table transformation formally. Then we describe the implemented facilities for handling summary tables. In the present study, we did not treat time-oriented data formally, but the function for handling time-oriented data has been implemented as one of the system utilities.

Table 1 Number of health personnel classified by organizations running the hospitals (1993)

Organizations ORG	doctors #DOCT	dentists #DENT	nurses #NURS	...
Ministry of Health and Welfare	6565	107	28258	...
other national	18993	2898	25664	...
cities,towns,villages	17606	399	61189	...
.	.	.	.	...
.	.	.	.	...
.	.	.	.	...

SUMMARY TABLES

Table 1 shows the numbers of doctors (denoted as #DOCT), dentists, etc. in hospitals classified by the organizations running hospitals (ORG). Table 2 shows the number of hospitals (#HOSP) classified by ORG and by the bed-size (BED). These are typical tables found in a published report of a national survey on clinical institutions. The attributes used

Table 2 Number of hospitals (#HOSP) classified by organizations running them and by bed-size (1993)

Bed-size (BED) Organizations (ORG)	20-29	30-39	40-49	50-99	...
Ministry of Health and Welfare	0	0	0	1	...
Ministry of Education	0	0	13	4	...
Labor Welfare Corporation	0	0	0	0	...
other national	1	3	0	15	...
cities,towns,villages	17	29	33	184	...
.	.	.	.	.	...
.	.	.	.	.	...
.	.	.	.	.	...

Table 3 Representation of Table 2 as a relation

ORG	BED	#HOSP
Ministry of Health and Welfare	20-29	0
Ministry of Health and Welfare	30-39	0
Ministry of Health and Welfare	40-49	0
.	.	.
.	.	.
.	.	.
Ministry of Education	20-29	0
Ministry of Education	30-39	0
Ministry of Education	40-49	13
.	.	.
.	.	.
.	.	.

to classify individuals such as ORG and BED are called category attributes. The attributes which describe some numerical properties of the groups of individuals such as #DOCT and #HOSP are called summary attributes. In publications, tables appear in various forms. For example, hospitals are classified in both rows and columns in Table 2. We place the basis of our study on the relational model, and a table like Table 2 is transformed into a relation (Table 3) where the categories of columns in Table 2 are represented as values of the category attribute BED. Hereafter, a table is assumed to be in the form of a relation where individuals are classified only in rows. There are also tables which show transformed values such as averages, percentages, etc. We assume that a table stores the counts (frequencies) or sums of summary attributes. Averages, percentages, etc., are treated as computed summary attributes. A computed summary attribute is defined by a numeric expression, e.g., if S1,S2 are summary attributes, S3=S1/S2 defines the computed summary attribute S3.

TABLE SCHEME

We will distinguish "a table occurrence" and "a table type." A table occurrence refers to an actual summary table, and by "a table," we will mean a table occurrence unless otherwise stated. A table type refers to the construct of a table, i.e., what attributes comprise the table and what values each attribute may take. Let T denote a table type. For a category attribute A of T, the set of categories from which A may take a value is called the domain of A in T and is denoted as $dom_T(A)$. If A is the sole category attribute in T, the set of rows is defined by $dom_T(A)$. If there are two category attributes A_1 and A_2 in T, the set of rows is defined by the Cartesian product $dom_T(A_1) \times dom_T(A_2)$ which is the set consisting of the ordered pairs (a,b) for every element a in $dom_T(A_1)$ and b in $dom_T(A_2)$. The set of rows is considered to be the same if the order of A_1 and A_2 is reversed. In general, let $A_1,...,A_l$ be the category attributes of T. We write $A_1,...,A_l$ as a set $\mathbf{A}=\{A_1,...,A_l\}$ since the set of rows is considered to be the same for any ordering of $A_1,...,A_l$. Similarly, the summary attributes of T are written as $\mathbf{S}=\{S_1,...,S_m\}$. The construct of a table type T is completely specified by the set of category attributes **A**, their respective domains, and the set of summary attributes **S**. These components of T is denoted as $[\{A.dom_T(A)|A \in \mathbf{A}\}, \mathbf{S}]$ and is called the table scheme of T.

As an example, let T be the table type of Table 3. ORG and BED are the category attributes, #HOSP is the summary attribute, and we have **A**={ORG,BED} and **S**={#HOSP}. As to ORG, three classification systems (the broad, the intermediate level, and the detailed) are used in the publication of the Ministry of Health and Welfare. Let $\mathbf{O_1},\mathbf{O_2},\mathbf{O_3}$ denote the corresponding sets of categories. As to BED, three classification systems shown in Table 4 are used. Let $\mathbf{B_1},\mathbf{B_2},\mathbf{B_3}$ denote the corresponding sets of categories. Table 3 is classified according to $\mathbf{O_1}$ of ORG and $\mathbf{B_1}$ of BED, and the scheme of T is represented as $[\{ORG.\mathbf{O_1}, BED.\mathbf{B_1}\}, \{\#HOSP\}]$. An SDB is a set of tables such that the construct of each table can be specified completely by a table scheme.

SUMMARY TABLE TRANSFORMATION

Suppose we have Table 1 and Table 3 stored on a disk. They both have the attribute ORG but are classified according to different classification systems. Table transformation allows a user to retrieve more tables than are stored, e.g., a table which is the same as Table 3 except that ORG is classified broadly, or a table which shows the columns of Table 1 and Table 3 together. For a system to perform such operations, it is necessary for the system to understand what are represented by the tables. In the following, we define six types of table transformation and describe what kind of knowledge (metadata) the system should have.

Table 4 Three classification systems of bed-size (BED)

B1	B2	B3
20- 29	20- 49	20- 49
30- 39		
40- 49		
50- 99	50- 99	50- 99
100-149	100-149	100-299
150-199	150-199	
200-299	200-299	
300-399	300-	300-499
400-499		
500-599		500-
.		
.		
.		

(a) category attribute A_1

C_1	C_2	C_3
a_1	b_1	c_1
a_2	b_2	
a_3		
a_4	b_3	c_2
a_5		

(b) category attribute A_2

D_1	D_2	D_3
x_1	y_1	z_1
x_2	y_2	
x_3	y_3	z_2
x_4		z_3

Fig.1 Classification systems of category attribute A_1 and A_2.

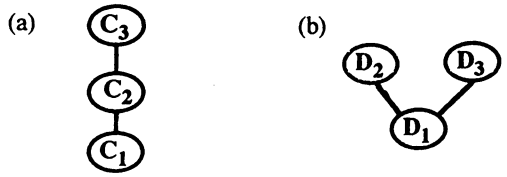

Fig.2 The Hasse diagrams of the relation $\leq$ on DOM(A_1) and DOM(A_2).

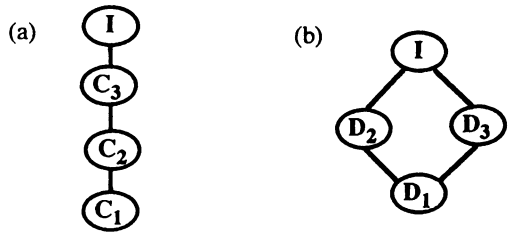

Fig.3 The Hasse diagrams of the relation $\leq$ on DOM(A_1) and DOM(A_2) after I is introduced.

Full Domains and Reducibility Relation

For a category attribute A, a set of categories given by a classification system will be called a full domain of A, and the set of full domains of A is denoted as DOM(A). As an illustrative example, let A_1 be a category attribute with three classification systems as shown in Fig.1 (a). A_1 is a simplified version of the attribute ORG. Let $\mathbf{C_1},\mathbf{C_2},\mathbf{C_3}$ be the corresponding sets of categories and we have DOM(A_1)= $\{\mathbf{C_1},\mathbf{C_2},\mathbf{C_3}\}$. As another example, let A_2 be an attribute with three classification systems as shown in Fig.1 (b). Let $\mathbf{D_1},\mathbf{D_2},\mathbf{D_3}$ be the corresponding sets of categories and we have DOM(A_2)=$\{\mathbf{D_1},\mathbf{D_2},\mathbf{D_3}\}$. A_2 is a simplified version of BED. There are tables which show only a part of categories given by some classification system. For a table type T and an attribute A, if $\mathrm{dom}_T(A)$ is a subset $\mathbf{C'}$ of a full domain $\mathbf{C}$, we write $\mathrm{dom}_T(A)=\mathbf{C}/\mathbf{C'}$ and call it a restricted domain. For example, let T be a table type classified according to $\mathbf{C_1}$ of A_1. If T shows only a_1,a_2,a_3,a_4 of $\mathbf{C_1}$, then $\mathrm{dom}_T(A_1)=\mathbf{C_1}/\{a_1,a_2,a_3,a_4\}$.

Now consider a request to obtain a table which is the same as Table 3 but with ORG classified broadly. Obviously, the request can be met by combining some rows of Table 3. But if a request is to change the classification of BED from $\mathbf{B_2}$ to $\mathbf{B_3}$ (Table 4), it cannot be met. To represent which domains can be converted to which domains for a given attribute A, we introduce a binary relation $\leq$ on DOM(A). For a full domain $\mathbf{C}$ of A, a partition of $\mathbf{C}$ is a set $\{\mathbf{P_1},\ldots,\mathbf{P_k}\}$ such that $\mathbf{P_i}$ is a subset of $\mathbf{C}$, $\cup\mathbf{P_i}=\mathbf{C}$, and $\mathbf{P_i}\cap\mathbf{P_j}=\phi$ for $i\neq j$. For full domains $\mathbf{C},\mathbf{C'}$ of A where $\mathbf{C'}=\{c_1,\ldots,c_k\}$, we define $\mathbf{C}\leq\mathbf{C'}$ if and only if there is a partition $\mathbf{P}=\{\mathbf{P_1},\ldots,\mathbf{P_k}\}$ of $\mathbf{C}$ such that $\mathbf{P_i}$ is a classification of c_i for $i=1,\ldots,k$. $\mathbf{P}$ is called a partition of $\mathbf{C}$ based on $\mathbf{C'}$. When $\mathbf{C}\leq\mathbf{C'}$ holds, $\mathbf{C}$ can be converted to (reduced to) $\mathbf{C'}$, and we call $\leq$ "the reducibility relation." The relation $\leq$ is a partial ordering on the set DOM(A) and it may be represented graphically by a Hasse diagram. Hasse diagrams of $\leq$ for A_1 and A_2 are shown in Fig.2. As to A_1, DOM(A_1)=$\{\mathbf{C_1},\mathbf{C_2},\mathbf{C_3}\}$, and $\mathbf{C_1}\leq\mathbf{C_2}$, $\mathbf{C_2}\leq\mathbf{C_3}$ and $\mathbf{C_1}\leq\mathbf{C_3}$ hold. A partition of $\mathbf{C_1}$ based on $\mathbf{C_2}$ $=\{b_1,b_2,b_3\}$ is $\mathbf{P}=\{\mathbf{P_1},\mathbf{P_2},\mathbf{P_3}\}$, where $\mathbf{P_1}=\{a_1\}$, $\mathbf{P_2}=\{a_2,a_3\}$, and $\mathbf{P_3}=\{a_4,a_5\}$. As to A_2, DOM(A_2) $=\{\mathbf{D_1},\mathbf{D_2},\mathbf{D_3}\}$, and $\mathbf{D_1}\leq\mathbf{D_2}$, $\mathbf{D_1}\leq\mathbf{D_3}$ hold.

Implicit Category Attributes

For a given survey, there is a population $\mathbf{P}$ (a set of individuals) on which the survey is conducted. For a given SDB, let $\mathbf{U}$ denote a set of every category attribute which appears in at least one table, and let $\mathbf{P}$ be the associated population. For every attribute A in $\mathbf{U}$, we introduce a conceptual full domain $\mathbf{I}$ into DOM(A) for the sake of consistency. $\mathbf{I}$ is defined to consist of a single category into which every individual taken from $\mathbf{P}$ falls, and hence $\mathbf{C}\leq\mathbf{I}$ holds for every element $\mathbf{C}$ in DOM(A). Fig.3 shows the Hasse diagrams for A_1 and A_2 after $\mathbf{I}$ is introduced. For a table type T with category attributes $\mathbf{A}$, we call any element B in $\mathbf{U}$-$\mathbf{A}$ an implicit category attribute and define that $\mathrm{dom}_T(B)=\mathbf{I}$, that is, every row has the same value (the sole element of $\mathbf{I}$) for B. We omit B from the scheme of T when $\mathrm{dom}_T(B)=\mathbf{I}$, so that if $\mathbf{A}$ is empty, T consists of a single row which represents the entire set of individuals. The table scheme of each table and the relation $\leq$ on the set DOM(A) for each attribute A in $\mathbf{U}$ comprise the metadata of the system.

Table Transformation

Domain Restriction. Let T be the table type of Table 1, and suppose it is requested to obtain a table which consists only of rows with the value of ORG "Ministry of Health and Welfare" or "other national." The required operation is to change dom_T(ORG) to $\mathbf{O_2}/\{$Ministry of Health and Welfare, other national$\}$, and we call this operation "domain restriction." In general, domain restriction selects a specified set of categories for a specified set of category attributes.

Domain Reduction. Let T be a table type classified according to $\mathbf{C_1}$ of A_1 (Fig.1 (a)). Suppose

it is requested to change the classification system from C_1 to C_2. We call the required operation "domain reduction." The definition is obvious when $dom_T(A_1)$ is a full domain. Now suppose $dom_T(A_1)$ is $C_1/\{a_1,a_2,a_3,a_4\}$. We have $C_1=\{a_1,a_2,a_3,a_4,a_5\}$, $C_2=\{b_1,b_2,b_3\}$, and b_1,b_2,b_3 correspond to $\{a_1\}$, $\{a_2,a_3\}$, and $\{a_4,a_5\}$ respectively. By domain reduction, a_4 is dropped and $dom_T(A_1)$ becomes $C_2/\{b_1,b_2\}$ since a_4 cannot be reduced to b_3 by itself. In general, for a table type T and an attribute A with $dom_T(A)=\mathbf{C/D}$, domain reduction on A is the operation of changing $dom_T(A)$ to $\mathbf{C'/D'}$ for such $\mathbf{C'}$ in DOM(A) that satisfies $\mathbf{C \le C'}$ where $\mathbf{D'}$ is given as follows: let $\mathbf{P}=\{\mathbf{P_1},...,\mathbf{P_k}\}$ be a partition of $\mathbf{C}$ based on $\mathbf{C'}=\{c_1,...,c_k\}$. $\mathbf{D'}$ is a set of such a category c_i in $\mathbf{C'}$ that the corresponding set $\mathbf{P_i}$ is included by $\mathbf{D}$.

Category Attribute Deletion. Suppose it is requested to obtain a table which is the same as Table 3 but is classified according to BED only. This is performed by deleting ORG from the scheme, and we call this operation "category attribute deletion." For a table type T with attributes $\mathbf{A}$, deletion of an attribute B is defined as follows. If $dom_T(B)$ is a full domain, B is simply deleted from $\mathbf{A}$, and in this case, deletion of B is equivalent to reducing $dom_T(B)$ to $\mathbf{I}$. When $dom_T(B)$ is $\mathbf{C/D}$ for a proper subset $\mathbf{D}$ of $\mathbf{C}$, $dom_T(B)$ cannot be reduced to $\mathbf{I}$. In this case, if there is a full domain $\mathbf{C'}$ that satisfies "$\mathbf{C \le C'}$ and $\mathbf{D}$ corresponds to a category d in $\mathbf{C'}$," $dom_T(B)$ is changed to $\mathbf{C'}/\{d\}$. If there is no such $\mathbf{C'}$, deletion of B is inhibited. The rationale for this is that deletion of B is equivalent to grouping all the categories in $dom_T(B)$ into a single category x which should represent some meaningful group by itself in the real world. So there should be a full domain to which x belongs. If deletion of B is necessary but is inhibited, the metadata should be modified. Deletion of multiple attributes is performed by repeatedly applying deletion of a single attribute.

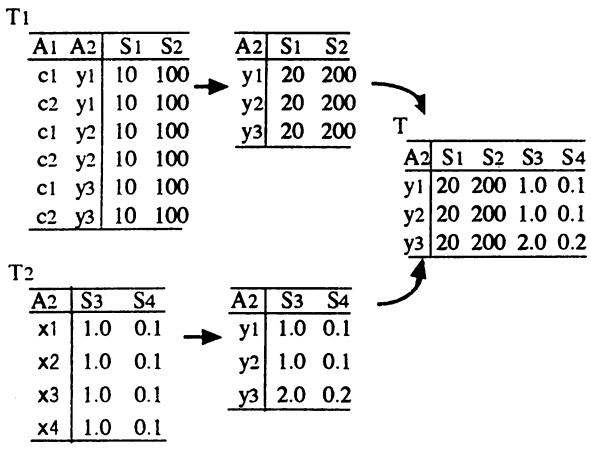

Fig.4 Table join of two table types T1 and T2.

Table Join. Fig.4 illustrates the operation of "table join." From two tables of type T_1 and T_2, where A_1 and A_2 are as given in Fig.1, the right most table of type T is produced. To join the two tables, the common set of rows is derived first by deleting A_1 from T_1 and reducing the domain of A_2 from $\mathbf{D_1}$ to $\mathbf{D_2}$ in T_2. Now we define join of tables of type T_1 and T_2, where $\mathbf{A_1}$, $\mathbf{A_2}$ and $\mathbf{S_1}$, $\mathbf{S_2}$ are their respective category attributes and summary attributes. The summary attributes of the resulting table type T is $\mathbf{S_1 \cup S_2}$. The category attributes $\mathbf{A}$ of T is obtained as follows. For A in $\mathbf{A_1 \cup A_2}$, let $\mathbf{C_1/D_1}$ and $\mathbf{C_2/D_2}$ be the domains of A in T_1 and T_2. $dom_T(A)$ should be chosen so that grouping of rows in T_1 and T_2 is kept as little as possible since the more rows are grouped, the more information is lost. To this end, the operation of "join" defined in algebraic structures is used. For $\mathbf{C_1}$ and $\mathbf{C_2}$ in DOM(A), "$\mathbf{C_1}$ join $\mathbf{C_2}$" is an element $\mathbf{C}$ in DOM(A) such that $\mathbf{C_1 \le C}$ and $\mathbf{C_2 \le C}$, and there exists no element $\mathbf{C'}$ in DOM(A) which satisfies $\mathbf{C_1 \le C' \le C}$ and $\mathbf{C_2 \le C' \le C}$. In the Hasse diagram, $\mathbf{C}$ is a node where ascending chains from $\mathbf{C_1}$ and $\mathbf{C_2}$ first join together. Let $\mathbf{C/D_1'}$ and $\mathbf{C/D_2'}$ be the reduced domains of $\mathbf{C_1/D_1}$ and $\mathbf{C_2/D_2}$ respectively, and $dom_T(A)$ is given as $\mathbf{C/(D_1' \cap D_2')}$. The set $\mathbf{A}$ is obtained by eliminating any attribute A with $dom_T(A)=\mathbf{I}$ from $\mathbf{A_1 \cup A_2}$.

Table Union and Summary Attribute Selection. Table union is an operation of putting the rows of two tables vertically. Summary attribute selection simply selects the specified summary attributes. The details are omitted here.

APPLICATION

A Sample SDB

An SDB on clinical institutions has been constructed as a sample database. The data source is a published report on a national survey on clinical institutions. We have selected tables to be stored by eliminating the redundant ones. When a table is organized as a combination of different table types, it is decomposed into multiple tables, each with its own table type. In all 180 tables were selected to comprise a database. We have no valid method for finding the minimal set of tables to be stored, and selection was made by inspection. In all 9 category attributes have been identified where the number of full domains varies from one to nine. The number of summary attributes is about 250. The system is being developed on a personal computer. As a relational database management system, ACCESS which supports SQL is used. For a given database, the metadata consists of a category attribute dictionary, a summary attribute dictionary, and a table scheme dictionary. The definition of a category attribute consists of the

definition of a set of full domains, the relation ≤ on the meta domain, and the sets of categories of each full domain. The definition of the relation ≤ includes the definition of join for every pair of full domains. The three dictionaries are actually stored as relations.

Operations on Summary Tables

Summary tables are retrieved basically by specifying the keywords. From the table types found, the user may select one or two types for further processing described below. Statistical analyses are left to application programs (such as statistical packages) outside the system.

Time-Oriented Information. Summary tables are obtained from surveys many of which are carried out periodically, i.e., annually, biannually, every five years, and so on. It is necessary for the system to be capable of handling time-oriented information. For this purpose, a time attribute is introduced. A time attribute is not an attribute that classifies a set of individuals observed at one time, but for each value of the attribute, the same set of individuals (may not coincide exactly) are observed repeatedly.

Table Transformation. The system maintains the schema of the transformed table types during the session, and the transformed tables may be stored if necessary. When new tables are stored, the table scheme dictionary is updated. A facility is provided for a database which involves a time attribute. In an SDB being constructed, a time attribute YEAR is introduced. For each table type, YEAR represents the years of the available table occurrences. For a given table type, a table transformation is performed on a set of table occurrences at one time. Table join and table union are performed only on those pairs of table occurrences which have the same value of YEAR.

Output Utilities. Any operations on tables besides table transformation are intended for output purposes. Output may be directed to a printer or a disk. By default, a table is produced in such a way that all the category attributes and summary attributes are arranged as columns side by side. Category attributes appear in the order as defined in a table scheme. The following may be performed optionally:

1. Specify the order of the category attributes.
2. Select one or two category attributes as column classifiers. Then one and only one summary attribute may be selected as table entries.
3. Define computed summary attributes.
4. For a given table type T with a time attribute H, the following tables may be produced:
   a) a table obtained by appending the specified table occurrences one after another.
   b) a table showing chronological transition of a specified summary attribute, where the rows are as defined by T and each column represents the summary attribute of each value of H.
   c) a table showing chronological transition of summary attributes for a specified row, where the rows correspond to the values of H and columns represent the summary attribute of T.
5. Produce a text file which is readily acceptable by the statistical software package SPSS.

DISCUSSION AND CONCLUSION

Various surveys are conducted in the field of health care. When one is interested in the survey results, it is not allowed to access the original data, and summary tables are the most detailed and hence the most important source of information which can be shared. To make the most of the summary tables, we have been developing an SDB management system. Based on the formal study, it was made clear what information should comprise the metadata and how it can be represented. Also the algorithms for handling tables could be implemented without ambiguity. Temporal information management has been studied deeply in the field of databases[5]. In our present study, we did not incorporate a time attribute into our formalism, but implemented the utilities for handling time-oriented information since the complexity involved in the discussion does not seem to be worthwhile in view of our study objective. The utilities for time-oriented information are not flexible enough and they should be more sophisticated in a future system extension. The method presented should contribute greatly to the health care filed where it is of urgent importance to put voluminous statistical data into organized databases so that they can be shared among people in the field.

References

1. Chan P, Shoshani A. SUBJECT: A Directory driven system for organizing and accessing large statistical databases. *Proc. 7th Int. Conf. Very Large Data Bases*, 1981: 553-563.
2. Sato H. Statistical Data Models: from a statistical table to a conceptual approach. *Statistical and Scientific Databases*, Michalewicz Z. Ed., Ellis Horwood, New York, 1991:167-200.
3. Catarci T, Santucci G. GRASP: a graphical system for statistical databases. *Proc. 5th Int. Conf. Statistical and Scientific Database Management*, 1990:148-162.
4. Ferri F, Grifoni P, Meo-Evoli L, Pisanelli DM, Ricci FL. ADAMS: Aggregate data management system for epidemiologists and health-care managers. *Computer Methods and Programs in Biomedicine*, 1993:40:43-53.
5. Shoshani A, Kawagoe K. Temporal data management. *Proc. 12th Int. Conf. Very Large Data Bases*, 1986:79-88.